Supplemental to Methods

Gene expression data analysis. Tools in *R* (1) and *Bioconductor* (2) were adapted for statistical analysis. Probe set summary measures were computed by robust multiarray averaging (3) applied to the combined set of 41 (31 tumor and 10 normal) arrays. To reveal global characteristics of the high-dimensional expression profiles, Euclidian distances were computed between every pair of profiles and used in hierarchical clustering and multi-dimensional scaling analyses. To identify genes with altered expression between tumor and normal tissues, several filters were applied: (1) gene-specific t-test controlling false discovery rate (FDR) at 0.1% using the q-value method (4), (2) empirical Bayes mixture modeling controlling FDR at 0.1% using the Gamma-Gamma model (5) and (3) consistent fold change of at least 1.5-fold in the same direction for all of 4 tumor-normal matched sample pairs. Other filters (e.g. moderated t statistics) had little additional effect and were not used in the final analysis. The two stringent statistical filters ensured high confidence in the final list of differentially expressed genes. The fold-change filter contributed information from the matched pair samples to eliminate patient to patient variability and ensure focus on genes with clear biological significance in NPC.

Spearman rank correlations measured associations between host and viral expression levels in the tumor samples. For each host probe set *h* and each viral gene *v*, the Spearman correlation *r(h,v)* was obtained by ranking the tumors by expression levels of *h* or *v*, and then by computing the Pearson correlation from these rank vectors. A set of *m=54,675* such correlations was computed for all viral measurements. Properties in these correlation sets should reflect significant association between host and virus expression at the whole profile level. For instance, with no association between host and virus expression, the distribution of *r(h,v)* ought to be centered at the value of 0 correlation. Four set features were considered for each viral gene: (1) minimum correlation, (2) maximum correlation, (3) median correlation, and (4) proportion of host genes with negative correlation with *v*. A permutation analysis assessed the significance of host-virus association as measured by these four set features. Sets of Spearman correlations were repeatedly derived (10,000 times) from host-virus expression data with randomly shuffled connecting tumor labels. Set features were computed each time; p-values measuring association were computed as the proportion of times in which the randomized feature was more extreme than the feature computed on the correct tumor labeling.

Gene Ontology (GO) Analysis. In both the tumor-normal comparison and the host-virus association study, host gene sets defined by specific GO annotations (6) were identified that were enriched for genes with altered expression. In total, 2354 different annotations associated with at least 10 probe sets on the Affymetrix microarray were scored for enrichment using the Pearson chi-square test for independence between annotation assignment and presence on the differential expression list (7). For the host-virus association study, a variation of this test was also used to improve sensitivity. Briefly, each host gene *g* had a numerical score *s(g)* measuring either differential expression or association with viral expression. Gene set *A* containing *n* genes received a raw enrichment score $X = \sum_{g \in A} s(g)$ and a standardized score $Z = (X - \mu)/\sigma$, where

$\mu = n\left(\sum_{g=1}^{m} s(g) \Big/ m\right)$ and $\sigma^2 = n\left(\frac{m-n}{m-1}\right)\left\{\left(\sum_{g=1}^{m} s^2(g) \Big/ m\right) - \left(\sum_{g=1}^{m} s(g) \Big/ m\right)^2\right\}$ were the mean and variance of X under random sampling of size *n* subsets from *m=54,675* probe sets, conditional upon the full set of gene-level scores *s(g)*. This general class of conditional tests improves power for detecting gene set enrichment (manuscript in preparation). Note that if scores *s(g)* are binary indicators of whether or not gene *g* is placed on a differential expression list, then X is the number of genes both on the list and in the gene set A, and the standardized Z^2 equals the Pearson chi-square statistic, up to the negligible factor $(m-1)/m$. We used both binary scores and scores based on the Spearman correlation between viral and host expression. To stabilize variance in the latter case, gene-level scores were computed as the inverse hyperbolic tangent of the correlation. Results are presented as Pearson-related Z scores (Z_b) and correlation-related Z scores (Z_q). Gene sets with the most extreme Z-scores were considered enriched for altered genes.

References

1. Ihaka R, Gentleman R. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 1996;5:299-314.
2. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80.
3. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4(2):249-64.
4. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100(16):9440-5.

5. Kendziorski CM, Newton MA, Lan H, Gould MN. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Stat Med 2003;22(24):3899-914.

6. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25(1):25-9.

7. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene expression. Genomics 2003;81(2):98-104.